Рецензия на доклад

Вероятностное тематическое моделирование и нейросетевые модели языка для обучения векторного представления слов, контекстов и документов

Доклад посвящен моделям представления текстовой информации с целью дальнейшего анализа. Докладчик описывает два основных направления исследования в этой области и представляет обзор различных моделей обоих типов, выявляя по ходу доклада их особенности и преимущества.

Среди прочих, рассматриваются две вероятностные модели для подсчета вероятностей слов в текстах: PLSA, вычисляющий вероятность слова в документе, и Skip-gram, где подсчитывается вероятность слова в некотором контексте. Выясняется, что выражения логарифмов правдоподобия в обоих случаях имеют схожее представление. Возникает серьезное подозрение о сводимости одной задачи к другой. Докладчик предлагает ввести в рассмотрение псевдо-документы: каждый псевдо-документ порождается словом и определяется, как мешок слов, встречающихся в локальном контексте любого вхождения этого слова. Теперь предсказание вероятностей слов в контекстах есть то же, что и обучение тематической модели на построенных псевдо-документах.

Далее докладчик формулирует задачу низкорангового матричного разложения для приближения частотного распределения слов в документе. Приводится ряд оптимизационных задач для целевых функций различных методов представления слов (PLSA, GloVe, SGNS), для каждой предъявлен оптимум с используемым методом оптимизации. Докладчик ставит вопрос о применимости вероятностных моделей к матрице слова-слова (с вероятностями встретить слово в окрестности другого слова). Представлены результаты экспериментов, согласно которым PLSA на такой матрице дает, с одной стороны, хорошее качество, а с другой — справляется с задачей аналогий: в построенном таким образом векторном пространстве выполняются не только естественные ограничения на близость (в смысле скалярного расстояния) близких по смыслу слов, но и некоторые семантические ограничения более высокого порядка (так, вектор log(«король») - log(«мужчина») + log(«женщина») оказывается близким к вектору log(«королева»)). Представлены экспериментальные результаты, демонстрирующие такого рода связи.

В заключительной части доклада представлены результаты построения тематической модели британского искусства на основе ручных описаний картин. Результаты позволяют, в частности, выделить основные темы и направления, характерные для британского искусства в определенный временной промежуток. Проведено ручное оценивание полученных тем для матрицы слова-слова и матрицы слова-документы. Приведенные цифры позволяют сделать вывод, что первый из этих подходов в целом справился с задачей лучше.

Хорошо проработанный материал и явная увлеченность докладчика темой сделали выступление живым и интересным. Данная область исследования представляется мне весьма перспективной и полезной с точки зрения общественного блага (очевидно, при надлежащем уровне успеха, подобная работа способна серьезно продвинуть, к примеру, область поиска информации). Наконец, серьезное впечатление производят достижения докладчика за предшествующий год и планы на будущий.